JOIV : Int. J. Inform. Visualization, 7(3-2): Empowering the Future: The Role of Information Technology in Building Resilience - November 2023 2040-2047

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Predicting Student's Soft Skills Based on Socio-Economical Factors: An Educational Data Mining Approach

Rathimala Kannan [a], Chew Chin Jet [b], Kannan Ramakrishnan [c,*], and Sujatha Ramdass [d]

[a] *PSG Institute of Management, PSG College of Technology, Coimbatore, India*
[b] *Faculty of Management, Multimedia University, Cyberjaya, Selangor, Malaysia*
[c] *Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia*
[d] *PSG Institute of Management, PSG College of Technology, Coimbatore, Tamil Nādu, India*
*Corresponding author: *kannan.ramakrishnan@mmu.edu.my*

Abstract*— Recent changes in the labor market and higher education sector have made graduates' employability a priority for researchers, governments, and employers in developed and emerging nations. There is, however, still a dearth of study about whether graduate students acquire the employability skills that businesses want of them because of their higher education. To determine a student's future employment and career path, it is critical to evaluate their soft skills. An emerging area called educational data mining (EDM) aims to gather enormous volumes of academic data produced and maintained by educational institutions and to derive explicit and specific information from it. This paper aims to predict students' soft skills such as professional, analytical, linguistic, communication, and ethical skills, based on their socio-economic, academic, and institutional data by leveraging data mining methods and machine learning techniques. All five soft skills were predicted using prediction models created using linear regression, probabilistic neural networks, and simple regression tree techniques. This study used a dataset from an open source that Universidad Technologica de Bolivar published. It covers academic, social, and economic data for 12,411 students. The experimental results demonstrated that the linear regression algorithm performed better than the others in predicting all five soft skills compared to machine learning methods. This finding can assist higher education institutions in making informed decisions, providing tailored support, enhancing student success and employability, and continuously modifying their programs to meet the needs of students.

***Keywords*— Prediction; machine learning; regression models; soft skills; higher education institutes.**

Manuscript received 10 Jan. 2023; revised 18 Jun. 2023; accepted 7 Sep. 2023. Date of publication 30 Nov. 2023.

I. INTRODUCTION

Educational Data Mining (EDM) is an emerging field focusing on collecting massive amounts of academic data generated and stored by educational institutions and extracting explicit and unique knowledge from the education sector. EDM is defined as performing data mining on educational data to predict trends and hidden patterns, which can be useful for data-driven decision-making [1]. By exposing hidden trends in academic achievement, machine learning algorithms assist decision-makers in developing plans to enhance students' learning experiences and institutional effectiveness [2]. EDM is used to make predictions of the student's performance in the future. It can identify the characteristics to improve the student's performance [3], [4]. In addition, EDM can predict the student's status, such as dropping out of school[5], finding out students who are losing interest in the course or subject as well as identifying slow learners [6], [7]. EDM can also provide feedback, support personalized learning, and help plan educational programs [8]–[10].

A. *Students' Soft Skills*

Students' Academic performance has been researched to reduce dropout rates, low academic achievement, and delayed graduation [1]. In recent years, the amount of data generated in the education sector has been extremely large. Data such as the coursework and test results can be analyzed to extract meaningful insight by implementing data mining techniques. The cumulative grade point average (CGPA) is the most popular measure of academic success [11], followed by coursework marks [12]. Therefore, universities should constantly monitor and analyze students' academic performance to alert and support students performing poorly or experiencing declining performance [13].

In today's dynamic, demanding, and international workplace, soft skills are a key predictor of graduate employment. Because of this and the rising expectations of employers, there is a growing need for a greater understanding of the significance of soft skills for graduates. Universities put a lot of effort into developing their students' soft skills [14]. A person's various learning, thinking, and acting styles are referred to as soft skills, which are dynamic, interpersonal psychological characteristics. Soft skills are known as "psychological attributes" because they can also be psychological characteristics that characterize a person's manner of being, unlike cognitive aptitudes [15]. Personality qualities influence soft skills, which impact the psychological processes that shape a person's vital development. Soft skills do not have a single definition. Most publications discuss soft skills as interpersonal, analytical, teamwork, communication, and collaborative skills [16]. Professional skills refer to the student's employability and career readiness. It enables students to transition from college to career and ensures that employers are satisfied with the students they are hiring [17]. A study conducted in New Zealand investigated soft skills required for software professionals and found that communication-related skills are most in demand in job advertisements [16]. Students always prefer analytical skills. A recent study [18] analyzed 3.8 million linkedIn job postings and found that analytical skills emerged as an important skill. In this study, "soft skills" refers to professional, analytical, linguistic, communication, and ethical skills.

B. Educational Data Mining

Educational Data Mining (EDM) is a rapidly evolving field that focuses on applying data mining techniques to educational settings. Researchers, practitioners, and a variety of stakeholders have been paying close attention to recent advancements in these fields because they are interested in learning more about how to use these data-driven technologies to improve student performance, enhance learning, and address potential issues in higher education like dropout rates, student retention and employability [19]. Literature reviews reveal many different applications of EDM, and the most common application is student performance prediction. Scholars have used most of the time cumulative grade point average (CGPA) as the measure of academic success [11], [19], [20] followed by coursework marks [12], [20].

Various types of variables are used as predictors of academic performance. A recent study from India predicted academic performance using demographic online usage data such as how often students checked their announcements, raised their hands during the online class, and the number of times they accessed the resources [21]. They have explored logistic regression and k-nearest neighbor algorithms and found logistic regression can predict better with a F1 score of 0.615. Another study assessed how well different machine learning methods, including Light Gradient Boosting Machine, Gradient Boosting, AdaBoost, Logistic Regression, Random Forest, and K-nearest Neighbors, predicted student academic performance. Along with student demographic factors, general grades, the family environment, and classroom interactions, they also considered learning styles and rates to classify whether a student has academic performance issues or not. Their study found that the gradient boosting algorithm outperformed with a F1 score of 0.9754 [2]. A survey of the literature indicated a dearth of studies on soft skills like professional, analytical, linguistics, communication, and ethical skills. Additionally, to our knowledge, no research has attempted to predict these soft skills using data mining and machine learning techniques.

Researchers, governments, and employers in developed and emerging economies now prioritize graduates' employability due to recent developments in the labor market and higher education sector. However, there is still a lack of research on the issue of whether graduate students receive the employability skills that companies demand from them because of their higher education [22]. People's learning, thinking, and acting processes are influenced by psychological traits known as soft skills. It is crucial to assess soft skills because they are indicators of future employment and career direction for students [15]. This paper aims to predict student's soft skills based on their socio-economical, academic, and institutional data by leveraging data mining methods and machine learning techniques [23], [24]. To address this problem, the following research questions were developed.

- RQ1. How are social and economic factors related to the student's soft skills?
- RQ2. How and to what extent social and economic factors can predict a student's soft skills, such as professional, analytical, linguistic, communication, and ethical skills?

This paper has contributed by comparing various machine learning algorithms' performances to predict student's soft skills such as professional, analytical, linguistics, communication, and ethical skills. Three regression algorithms, linear regression, probabilistic neural network (PNN), and simple regression tree, were applied to build the machine learning models for each of the five soft skills. Most research has employed student-related data to predict students' academic achievement, but there is comparatively little literature available on teacher- or institution-related factors [19]. This study advances knowledge by examining the role of educational institutions' socio-economic level in predicting students' soft skills.

II. Materials and Method

The Cross-Industry Standard Process-Data Mining (CRISP-DM) framework was used in this study to develop prediction models. CRISP-DM comprises six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [25]. The research problem and data mining objectives are clearly identified in the business understanding phase. The second phase focuses on data collection and explorative data analysis. The data was obtained from an open source, and the initial descriptive analytics were carried out. The next phase is data preparation where the raw data was converted into structured data by removing redundant rows and columns that are irrelevant to the problem. Missing values are imputed with median values for integer values and the most frequent values for string values. After preprocessing, data was partitioned into training, validation, and test datasets.

The training and validation dataset was used to train the machine learning model and validation in the modeling phase. Three regression models were developed by applying simple linear regression, PNN, and simple regression tree algorithms. Hyperparameters were optimized to identify the optimal models. The next phase is the evaluation of machine learning models using the test dataset. Here, the testing performance of each regression model was compared, and the best model was identified. The last phase is the deployment, where the best model is recommended. Figure 1 shows the activities that were done during each of these phases for this research.

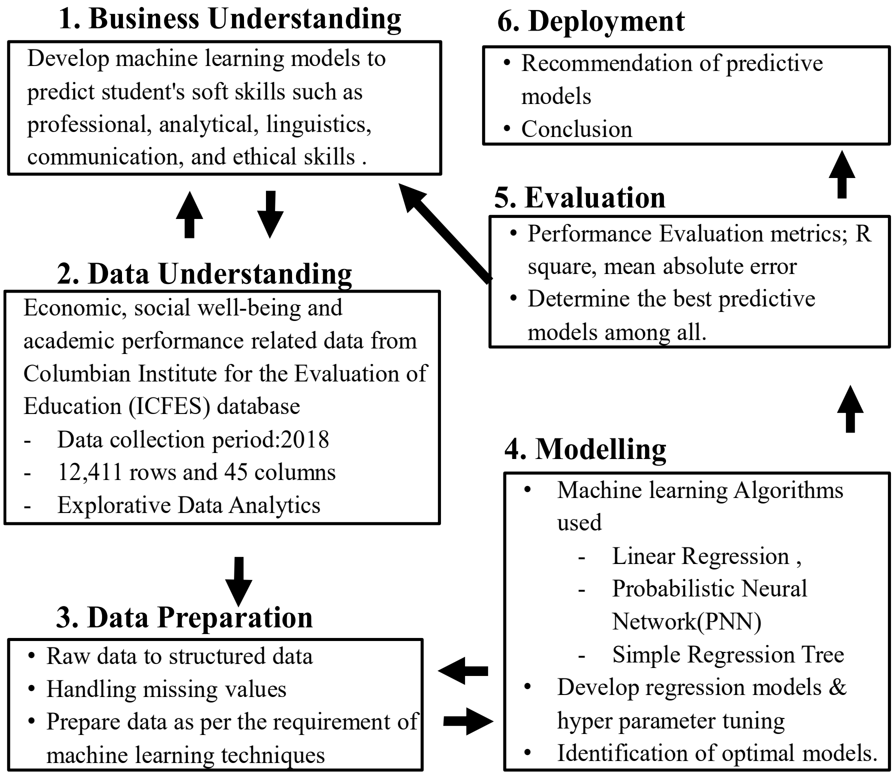

Fig. 1 Research Methodology

A. Data Collection

This study adopted a dataset from an open source published by Universidad technologica de bolivar under education, educational assessment, academic learning, and predictive modeling. The Colombian Institute for Education Evaluation data have been obtained [26], [27]. The data represents the national assessments for secondary and university education in engineering students. It contains academic, social, and economic information for 12,411 students. More detailed information about the dataset can be obtained from [28]. The dataset consists of 11 demographic variables, including gender, student's education, parents' occupation, school, and university information. There were 16 socio-economic variables such as resident status, economic aid provided by the government, socio-economic level of the family, number of people in the household, job, income, access to the internet, TV, computer, washing machine, microwave oven, car, DVD, phone, and mobile. Also, the dataset consists of 15 academic-related variables, such as marks obtained for different subjects from school and university.

B. Data Preparation

Several activities were carried out to prepare the data suitable to proceed with modeling. The original dataset had some values in Spanish, which were translated to English. Some variables had inconsistent values, which were identified and corrected. Also, some variables are encoded from string to integer. Some variables, such as percentile, 2nd_Decile, and quartile, were highly correlated with the global score variable; therefore, they are removed from further analysis. Since the original dataset did not have soft skills-related variables, they are computed from existing subjects' marks. The table below depicts how the five soft skills variables are computed. Skills derived from more than one subject used the average marks.

TABLE I
COMPUTATION OF SOFT SKILLS

Skill Name	Test(s) used
Professional Skills	Formulation of Engineering Project (FEP_PRO)
Analytical Skills	Mathematics (MAT_S11) & Quantitative Reasoning (QR_PRO)
Linguistic Skills	English (ENG_S11) & English (ENG_PRO)
Communication Skills	Critical Reading (CR_S11) & Critical Reading (CR_PRO) & Written Communication (WC_PRO)
Ethical Skills	Citizen Competencies (CC_S11) & Citizen Competencies (CC_PRO)

*Note: _S11 refers to the Saber 11 test, which is the high school test, and _PRO refers to the Saber Pro test, which is the university test.

C. Machine Learning Models

This study compares the performance of various machine learning techniques such as logistic regression, probabilistic neural networks, and simple regression tree–regression model from decision tree algorithm [24] to predict student's various soft skills such as professional skills, analytical skills, linguistic skills, communication skills, and ethical skills. Machine learning models from these three algorithms are trained, validated, and evaluated to predict each soft skill. During the modeling phase, the hyperparameters of each model were tuned to identify the optimal setting to identify the optimal model. Standard performance measures of regression models such as R^2, and mean absolute error were used to identify the best prediction model. R^2 is frequently used in regression analysis to gauge how well a model fits the data. It has a scale of 0 to 1, with 0 denoting that the model does not explain any variance in the dependent variable, 1 means that the model predicts the dependent variable with absolute accuracy.

Mean Absolute Error (MAE) is a metric used to gauge the average size of discrepancies between actual and predicted values in a regression problem. It indicates how similar, on average, forecasts are to the actual results [29]. MAE is frequently used in statistics and machine learning to assess the effectiveness of regression models. It is especially helpful when there are outliers in the data because it penalizes huge errors less severely than other metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). MAE offers a simple explanation as the average absolute departure from the genuine values[30]. When actual values (ri) and predicted (pi) values are given, the metrics are calculated as shown below.

- $R^2 = 1 - SS_{res} / SS_{tot} = 1 - \Sigma(p_i - r_i)^2 / \Sigma(r_i - 1/n*\Sigma r_i)^2$ (Can be negative!),
- Mean absolute error ($1/n*\Sigma|p_i - r_i|$),
- Mean squared error ($1/n*\Sigma(p_i - r_i)^2$),
- Root mean squared error ($\mathrm{sqrt}(1/n*\Sigma(p_i - r_i)^2)$)

III. Results and Discussion

A. How are Social and Economic Factors Related to the Student's Soft Skills?

Social and economic circumstances can greatly impact the development of a student's professional skills. These variables may influence a student's experiences, opportunities, and resources, which may eventually affect their capacity to develop and improve their professional abilities. The findings of this study demonstrated a strong relationship between students' socio-economic status and their level of professional skill. Figure 2 shows that the median student professional scores for socio-economic levels 1 to 4 are 149, 151, 153, and 160. However, the socio-economic standing of the higher education institution and the student's professional skills do not correlate linearly. Most students enroll in level 3 higher education institutions; their median professional competence score is the lowest. However, students studying from the higher socio-economic level institutions exhibited the highest median professional competence.

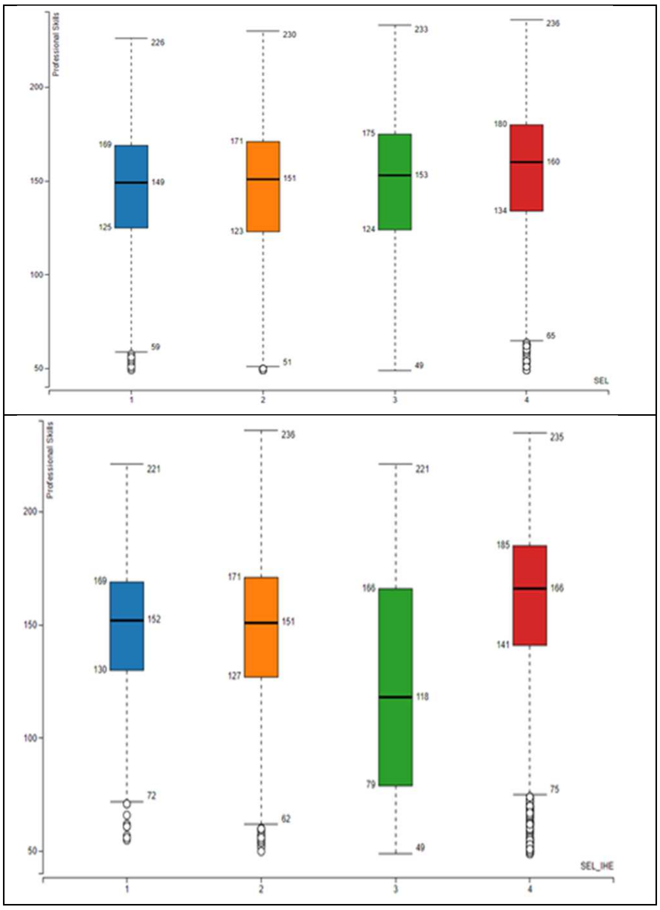

Fig. 2 Professional skills vs Socio-economic level of the student and the higher education institute

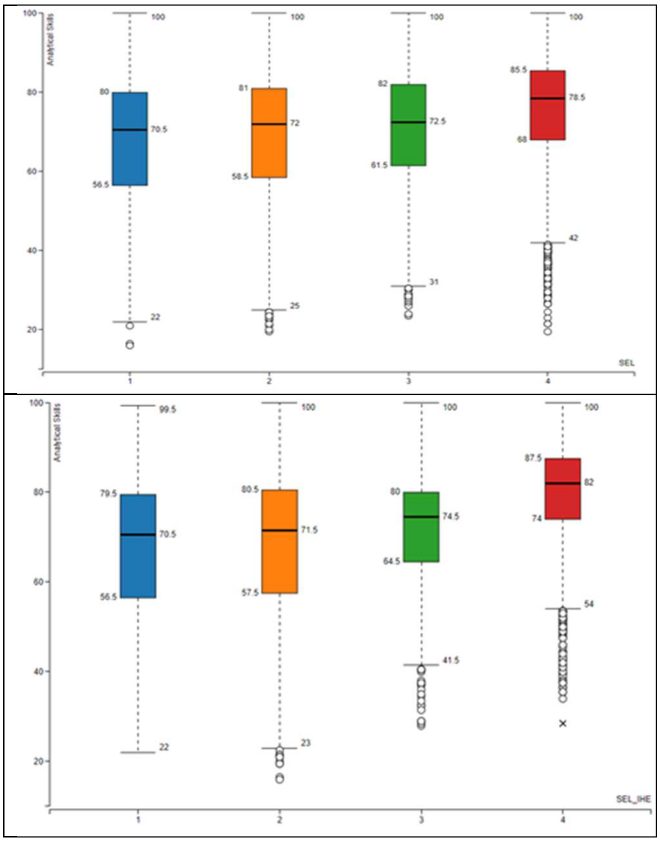

Fig. 3 Analytical skills vs socio-economic level of the student and the higher education institute

The development of analytical skills can be impacted by social and economic elements in a student's home environment. Students' ability to think critically and solve problems is more likely to develop in families that value education and create a conducive environment for learning. Figure 3 illustrates that students' analytical skills are positively related to their home's socio-economic level. Unlike professional skills, the results showed student's analytical skills are linearly related to their higher education institution's socio-economic level.

Financial factors can influence who has access to language resources and learning aids. Figure 4 depicts the relationship between linguistics skills and the socio-economic level of the student and the higher education institute. There is a noticeable difference in the student's linguistic skills among the different levels of their socio-economic status. The opportunity to acquire books, educational programs, and language-related technologies may be greater for students from higher-income families, which could help them improve their language abilities. Students are more likely to obtain thorough language education and improve their linguistic abilities if they attend schools with well-funded language programs and qualified language teachers.

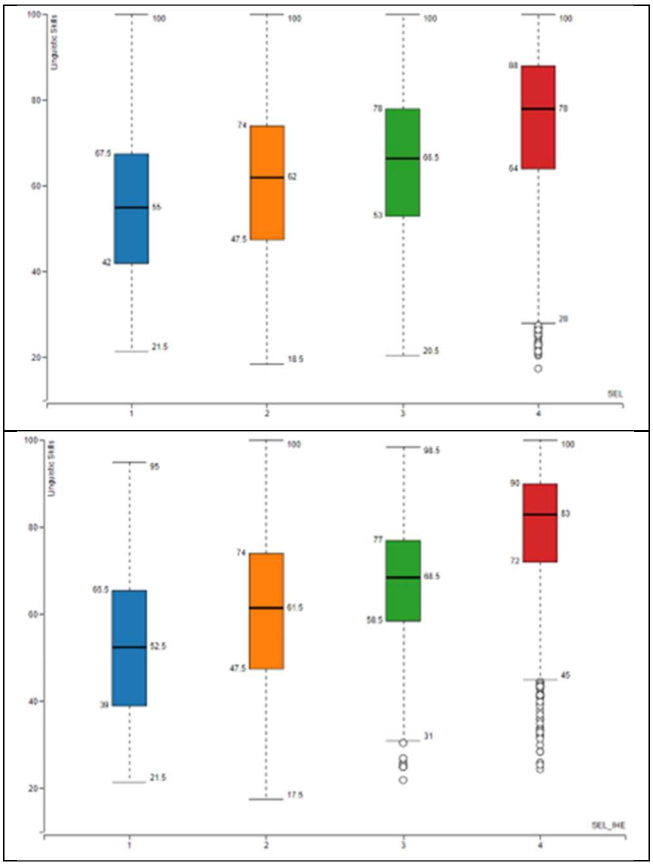

Fig. 4 Linguistics skills vs socio-economic level of the student and the higher education institute

Social and economic circumstances can impact a student's communication abilities, particularly parental education levels. Figure 5 presents the communication skills of students from different socio-economic levels. Higher-educated parents are more likely to foster a language-rich environment for their children, engage in effective communication strategies, and provide encouragement for developing their communication skills. Communication skills can be impacted by economic factors that limit a student's access to a high-quality education. Students from higher socio-economic origins have access to institutions with superior resources, knowledgeable instructors, and possibilities for communication-focused extracurriculars like presentations, debates, and public speaking. Effective communication abilities are more likely to develop in students who attend schools with well-funded communication programs and extra-curricular public speaking, debate, theatre, or writing activities.

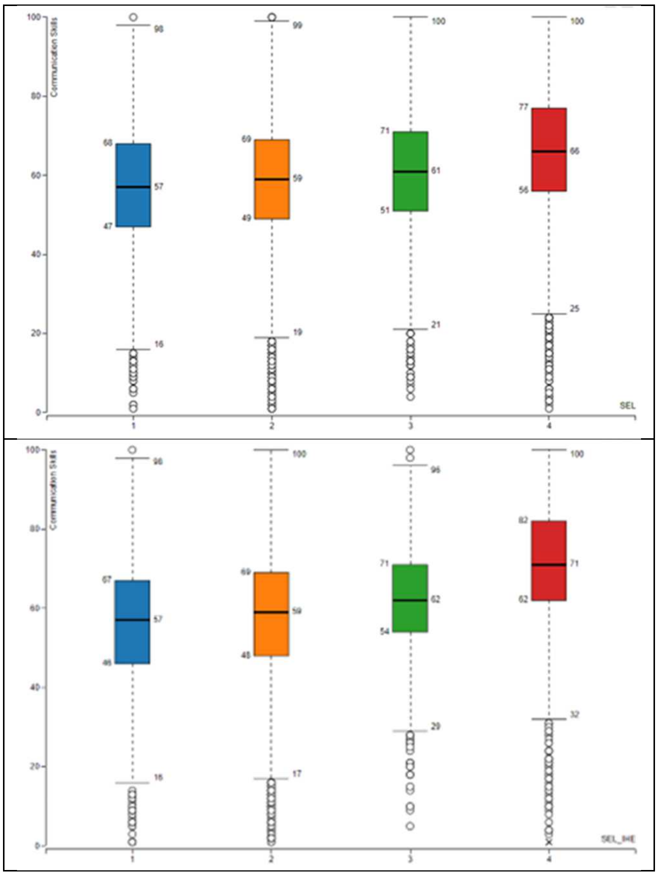

Fig. 5 Communication skills vs socio-economic level of the student and the higher education institute

Economic differences, such as poverty and wealth inequalities, can impact students' ethical perspectives and influence how they perceive social duty and fair play. Figure 6 presents the relationship between ethical skills and the socio-economic level of the student and the higher education institute. A student's ethical abilities are greatly influenced by social influences, particularly parental influence, irrespective of their socio-economic status. This is noticeable from Figure 6. Also, economic issues may impact the accessibility and caliber of educational materials and initiatives that foster ethical behavior. Students attending schools with robust moral education initiatives or moral curricula may have greater opportunities to hone their capacity for moral reflection, empathy, and judgment.

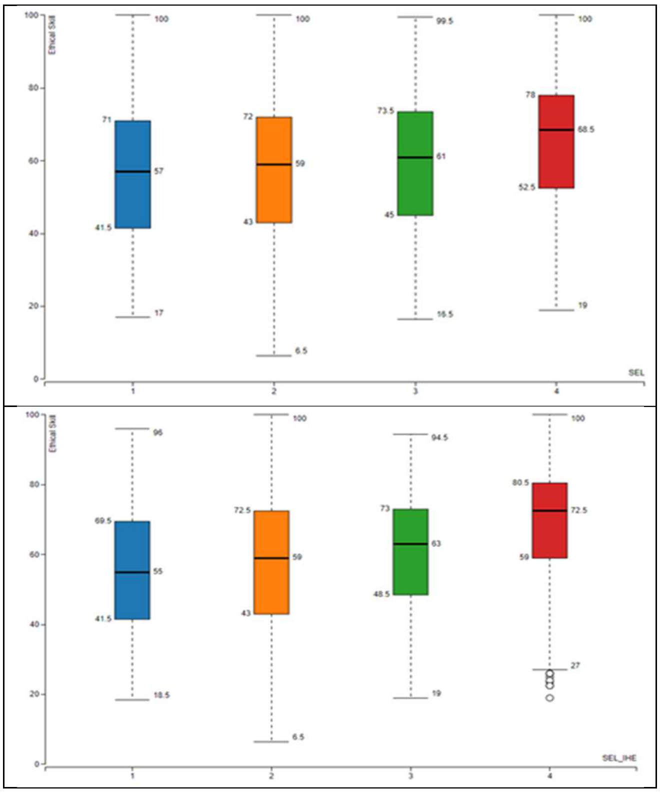

Fig. 6 Ethical skills vs Socio-economic level of the student and the higher education institute

B. How and to What Extent Social and Economic Factors can Predict a Student's Soft Skills such as Professional, Analytical, Linguistics, Communication, and Ethical Skills?

To answer this question, three regression algorithms, linear regression, probabilistic neural network (PNN), and simple regression tree, were applied to build the machine learning models for each of the five soft skills.

TABLE II
PROFESSIONAL SKILLS PREDICTION MODELS' TRAINING (MODELLING) PERFORMANCE

Regression Technique	Modeling: Evaluation	Modeling Training: Validation	R^2 value	MAE	MSE	RMSE
Linear Regression Model	**60:40**	**70:30**	**0.22**	**0.133**	**0.03**	**0.173**
PNN Model	60:40	50:50	-0.482	0.182	0.056	0.236
Simple Regression Tree Model	60:40	50:50	0.016	0.15	0.038	0.195

TABLE III
ANALYTICAL SKILLS PREDICTION MODELS' TRAINING (MODELLING) PERFORMANCE

Regression Technique	Modeling: Evaluation	Modelling Training: Validation	R^2 value	MAE	MSE	RMSE
Linear Regression Model	**50:50**	**60:40**	**0.367**	**0.115**	**0.022**	**0.149**
PNN Model	60:40	50:50	-0.307	0.163	0.046	0.213
Simple Regression Tree Model	50:50	70:30	-0.499	0.181	0.053	0.23

Table 2,3,4,5,6 compares the training performance of these three algorithms to predict professional, analytical, linguistic, communication, and ethical skills, respectively. The various partitioning ratios were identified as the optimal settings by tuning the hyperparameters of each algorithm. In the modeling phase, the results showed that the simple linear regression model outperformed the rest of the machine-learning techniques for all the soft skills investigated in this paper.

TABLE IV
LINGUISTIC L SKILLS PREDICTION MODELS' TRAINING (MODELLING) PERFORMANCE

Regression Technique	Modelling: Evaluation	Modelling Training: Validation	R^2 value	MAE	MSE	RMSE
Linear Regression Model	**50:50**	**70:30**	**0.656**	**0.104**	**0.018**	**0.133**
PNN Model	60:40	70:30	-0.023	0.188	0.057	0.239
Simple Regression Tree Model	50:50	50:50	0.315	0.147	0.035	0.187

TABLE V
COMMUNICATION SKILLS PREDICTION MODELS' TRAINING (MODELLING) PERFORMANCE

Regression Technique	Modelling: Evaluation	Modelling Training: Validation	R^2 value	MAE	MSE	RMSE
Linear Regression Model	**50:50**	**50:50**	**0.699**	**0.08**	**0.011**	**0.103**
PNN Model	70:30	70:30	-0.061	0.146	0.036	0.19
Simple Regression Tree Model	70:30	60:40	0.375	0.115	0.022	0.149

TABLE VI
ETHICAL SKILLS PREDICTION MODELS' TRAINING (MODELLING) PERFORMANCE

Regression Technique	Modelling: Evaluation	Modelling Training: Validation	R^2 value	MAE	MSE	RMSE
Linear Regression Model	**60:40**	**50:50**	**0.64**	**0.1**	**0.016**	**0.128**
PNN Model	60:40	60:40	-0.441	0.203	0.067	0.258
Simple Regression Tree Model	50:50	70:30	0.36	0.138	0.031	0.176

Following the modeling phase, all the optimal prediction models were evaluated using the test data. Table 7 presents the performance of all prediction models for each soft skill. Consistent with the modeling performance, linear regression models again performed the best among the three algorithms for all soft skills. The results show that social and economic factors can predict 59.5% of ethical skills, 49.5% of communication skills, 40.7% of linguistic skills, and 35.6% of analytical skills. However, social and economic factors cannot predict professional skills, with a lower predictability of 4.3%. Machine learning algorithms are used in educational data mining to predict students' future behaviors by first learning from their historical data. The reliability and relevance of the data have a major impact on the predicted performance [19]. Additionally, this study's findings show that traditional linear regression is superior to probabilistic neural networks and simple regression tree methods for predicting all soft skills.

TABLE VII
SUMMARY OF ALL PREDICTION MODELS TESTING (EVALUATION) PERFORMANCE

Skill Category	Regression Technique	R^2 value	MAE	MSE	RMSE
Professional Skills	**Linear Regression Model**	**0.043**	**0.135**	**0.034**	**0.185**
	PNN Model	-0.078	0.132	0.039	0.197
	Simple Regression Tree	-0.042	0.148	0.037	0.193
Analytical Skills	**Linear Regression Model**	**0.356**	**0.119**	**0.024**	**0.156**
	PNN Model	0.108	0.116	0.033	0.183
	Simple Regression Tree	-0.003	0.148	0.038	0.194
Linguistic Skills	**Linear Regression Model**	**0.407**	**0.14**	**0.031**	**0.177**
	PNN Model	0.401	0.109	0.032	0.178
	Simple Regression Tree	0.143	0.164	0.045	0.213
Communication Skills	**Linear Regression Model**	**0.495**	**0.1**	**0.017**	**0.13**
	PNN Model	0.463	0.073	0.018	0.134
	Simple Regression Tree	0.044	0.142	0.032	0.178
Ethical Skills	**Linear Regression Model**	**0.595**	**0.093**	**0.015**	**0.121**
	PNN Model	0.052	0.116	0.034	0.185
	Simple Regression Tree	0.536	0.083	0.017	0.129

Findings from this study show that promoting the equitable development of soft skills among students requires addressing social and economic inequalities. The emphasis should be on giving students from all socio-economic levels equal access to leadership opportunities, mentorship programs, and encouraging learning settings. It is possible to lessen the negative effects of these issues on soft skills development by creating inclusive learning settings that support cultural awareness, fostering soft skill development, and providing targeted support and resources for students who may experience social or economic difficulties.

Learning experiences can be adapted to meet the needs of each student by predicting their soft skills. To better assist students' overall development, educators can create individualized learning plans, offer relevant co-curricular activities, and provide focused support by knowing a student's soft skill strengths and areas for development. The importance of soft skills in influencing student achievement and retention is becoming more widely acknowledged. Institutions can identify struggling students by predicting and tracking their soft skills. They can then offer those students prompt interventions, support services, or mentorship programs. As a result, there may be an increase in academic achievement overall and in student involvement and happiness.

IV. Conclusion

This paper examined and compared the performance of three regression algorithms to predict students' soft skills, such as professional skills, analytical skills, linguistic skills, communication skills, and ethical skills, based on socioeconomic, academic, and institutional data of students from higher education institutions. The experiment outcomes demonstrated that machine learning algorithms can be applied to predict students' soft skills. Prediction models for each of the five soft skills were developed using traditional statistical methods like linear regression and machine learning techniques like probabilistic neural networks and simple regression trees.

The findings indicated that of the three methods based on R^2 and MAE, linear regression models surpassed them as the best for predicting all soft skills. The results show that social and economic factors can predict 59.5% of ethical skills, 49.5% of communication skills, 40.7% of linguistic skills, and 35.6% of analytical skills. However, social and economic factors cannot predict professional skills, with a lower predictability of 4.3%. Predictive analytics algorithms initially learn from their prior data to predict student's future behaviors. To make accurate predictions, it is crucial to collect trustworthy data. The future study could collect more data related to soft skills to have a better prediction model. In conclusion, predicting students' soft skills can help higher education institutions make wise choices, give individualized support, improve student success and employability, and continuously adapt their programs to match the needs of students.

References

[1] S. M. Dol and P. M. Jawandhiya, "Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey," *Eng Appl Artif Intell*, vol. 122, p. 106071, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106071.

[2] N. N. Sanchez-Pozo, J. S. Mejia-Ordonez, D. C. Chamorro, D. Mayorca-Torres, and D. H. Peluffo-Ordonez, "Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques," *Future of Educational Innovation Workshop Series - Machine Learning-Driven Digital Technologies for Educational Innovation Workshop 2021*, 2021, doi: 10.1109/IEEECONF53024.2021.9733756.

[3] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V Moniaga, "Determining factors that affect student performance using various machine learning methods," *Procedia Comput Sci*, vol. 216, pp. 597–603, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.174.

[4] R. Alcaraz, A. Martinez-Rodrigo, R. Zangroniz, and J. J. Rieta, "Early Prediction of Students at Risk of Failing a Face-to-Face Course in Power Electronic Systems," *IEEE Transactions on Learning Technologies*, vol. 14, no. 5, pp. 590–603, Oct. 2021, doi: 10.1109/TLT.2021.3118279.

[5] D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, and W. Verbeke, "Uplift Modeling for preventing student dropout in higher education," *Decis Support Syst*, vol. 134, p. 113320, Jul. 2020, doi: 10.1016/J.DSS.2020.113320.

[6] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, pp. 1–19, Dec. 2022, doi: 10.1186/S40561-022-00192-Z/TABLES/14.

[7] Y. Cheng, B. Pereira Nunes, and R. Manrique, "Not Another Hardcoded Solution to the Student Dropout Prediction Problem: A Novel Approach Using Genetic Algorithms for Feature Selection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13284 LNCS, pp. 238–251, 2022, doi: 10.1007/978-3-031-09680-8_23.

[8] S. Alturki, N. Alturki, and H. Stuckenschmidt, "Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions," *Journal of Information Technology Education: Innovations in Practice*, vol. 20, pp. 121–137, Jul. 2021, doi: 10.28945/4835.

[9] R. Kannan *et al.*, "Advancements in Machine Learning Techniques for Educational Data Mining: An Overview of Perspectives and Trends," *International Journal of Membrane Science and Technology*, vol. 10, no. 3, pp. 1820–1839, Sep. 2023, doi: 10.15379/IJMST.V10I3.1841.

[10] A. A. Al Hassani and S. Wilkins, "Student retention in higher education: the influences of organizational identification and institution reputation on student satisfaction and behaviors," *International Journal of Educational Management*, vol. 36, no. 6, pp. 1046–1064, Jan. 2022, doi: 10.1108/IJEM-03-2022-0123.

[11] S. Rajendran, S. Chamundeswari, and A. A. Sinha, "Predicting the academic performance of middle- and high-school students using machine learning algorithms," *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100357, Jan. 2022, doi: 10.1016/J.SSAHO.2022.100357.

[12] O. Ojajuni *et al.*, "Predicting Student Academic Performance Using Machine Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12957 LNCS, pp. 481–491, 2021, doi: 10.1007/978-3-030-87013-3_36.

[13] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a Web-Based Prediction System for Students’ Academic Performance," *Data 2022, Vol. 7, Page 21*, vol. 7, no. 2, p. 21, Jan. 2022, doi: 10.3390/DATA7020021.

[14] A. Asefer and Z. Abidin, "Soft Skills and Graduates' Employability In The 21st Century From Employers' Perspectives: A Review of Literature," *International Journal of Infrastructure Research and Management*, vol. 9, no. 2, pp. 44–59, 2021, [Online]. Available: https://iukl.edu.my/rmc/publications/ijirm/

[15] Á. Escolà-Gascón and J. Gallifa, "How to measure soft skills in the educational context: psychometric properties of the SKILLS-in-ONE questionnaire," *Studies in Educational Evaluation*, vol. 74, p. 101155, Sep. 2022, doi: 10.1016/J.STUEDUC.2022.101155.

[16] M. Galster, A. Mitrovic, S. Malinen, J. Holland, and P. Peiris, "Soft skills required from software professionals in New Zealand," *Inf Softw Technol*, vol. 160, p. 107232, Aug. 2023, doi: 10.1016/J.INFSOF.2023.107232.

[17] S. DeArmond, B. L. Rau, J. Buelow-Fischer, A. Desai, and A. J. Miller, "Teaching professional skills during the pandemic: Does delivery mode matter?," *The International Journal of Management Education*, vol. 21, no. 2, p. 100770, Jul. 2023, doi: 10.1016/J.IJME.2023.100770.

[18] K. M. Peesker, P. D. Kerr, W. Bolander, L. J. Ryals, J. A. Lister, and H. F. Dover, "Hiring for sales success: The emerging importance of salesperson analytical skills," *J Bus Res*, vol. 144, pp. 17–30, May 2022, doi: 10.1016/J.JBUSRES.2022.01.070.

[19] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)," *Educ Inf Technol (Dordr)*, pp. 1–35, Dec. 2022, doi: 10.1007/S10639-022-11536-0/Figures/1.

[20] T. Mansouri, A. Z., and A. Ashrafi, "A Learning Fuzzy Cognitive Map (LFCM) Approach to Predict Student Performance," *Journal of Information Technology Education: Research*, vol. 20, pp. 221–243, May 2021, doi: 10.28945/4760.

[21] N. Sharma and M. Yadav, "A Comparative Analysis of Students' Academic Performance using Prediction Algorithms Based on Their Time Spent on Extra-Curricular Activities," *Proceedings of the 2022 3rd International Conference on Intelligent Computing, Instrumentation and Control Technologies: Computational Intelligence for Smart Systems, ICICICT 2022*, pp. 745–750, 2022, doi: 10.1109/ICICICT54557.2022.9917606.

[22] O. B. O. Damoah, A. A. Peprah, and K. O. Brefo, "Does higher education equip graduate students with the employability skills employers require? The perceptions of employers in Ghana," *https://doi.org/10.1080/0309877X.2020.1860204*, vol. 45, no. 10, pp. 1311–1324, 2021, doi: 10.1080/0309877X.2020.1860204.

[23] S. Ranjeeth, T. P. Latchoumi, and P. V. Paul, "A Survey on Predictive Models of Learning Analytics," *Procedia Comput Sci*, vol. 167, pp. 37–46, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.180.

[24] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," *Applied Sciences 2020, Vol. 10, Page 1042*, vol. 10, no. 3, p. 1042, Feb. 2020, doi: 10.3390/APP10031042.

[25] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, Jan. 2021, doi: 10.1016/J.PROCS.2021.01.199.

[26] E. Delahoz-Dominguez, R. Zuluaga, and T. Fontalvo-Herrera, "Dataset of academic performance evolution for engineering students," *Data Brief*, vol. 30, p. 105537, 2020, doi: 10.1016/j.dib.2020.105537.

[27] R. Zuluaga-Ortiz, A. Camelo-Guarín, and E. Delahoz-Domínguez, "Efficiency analysis trees as a tool to analyze the quality of university education," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4412–4421, Aug. 2023, doi: 10.11591/IJECE.V13I4.PP4412-4421.

[28] E. De La Hoz, "Data of Academic Performance evolution for Engineering Students," vol. 1, 2020, doi: 10.17632/83TCX8PSXV.1.

[29] B. Panigrahi, K. C. R. Kathala, and M. Sujatha, "A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models," *Procedia Comput Sci*, vol. 218, pp. 2684–2693, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.241.

[30] J. Behera, A. K. Pasayat, H. Behera, and P. Kumar, "Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets," *Eng Appl Artif Intell*, vol. 120, p. 105843, Apr. 2023, doi: 10.1016/J.ENGAPPAI.2023.105843.